Make Chicago "Safe" Again

Data Mining Project

Luis Veltze
University of Colorado at Booulder
Boulder, Colorado
luis.veltze@colorado.edu

Ryan Close
University of Colorado at Boulder
Boulder, Colorado
ryan.close@colorado.edu

Tyler Mooney
University of Colorado at Boulder
Boulder, Colorado
tyler.mooney@colorado.edu

Garrett Glissmann
University of Colorado at Boulder
Boulder, Colorado
garrett.glissmann@colorado.edu

ABSTRACT

The goal of this project is to find interesting trends using Chicago's crime reports spanning the years 2001 to early 2018. We will perform a longitudinal survey of crime trends. We hope to find geographical patterns as well as discover connections to other public health factors and socioeconomic indicators. For example, we could correlate crime with birth rates or access to affordable housing and grocery stores.

ACM Reference format:
Luis Veltze, Ryan Close, Tyler Mooney, and Garrett Glissmann. 2016. Make Chicago "Safe" Again. In *Proceedings of CSCI-4502,*
, Boulder (University of Colorado at Boulder), 3 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 PROBLEM STATEMENT

In our search to find a interesting dataset that would be a good candidate for data mining, we considered datasets about bitcoin and crime. We selected crime data because it was more intuitive to understand the attributes and ultimately mine correlations between similar datasets. Crime data is one of many subjects that is continuously examined and analyzed for trends and patterns. Of course, studying crime has an important social value: to understand how much crime is happening, how it compares in terms of time and geography, and possible reasons that crime rises or falls. The hope is that in understanding the patterns and causes of crime, it might be possible to bring about changes that can reduce crime. This is evident in policing today in terms of directed patrols to crime "hotspots" within cities. Using the crime data and other public data recorded by the city of Chicago, this team will search for trends in crime that perhaps could be used for social action.

University of Colorado at Boulder,

DOI: 10.1145/nnnnnnn.nnnnnnn

2 LITERATURE SURVEY

The existing content analyzed in this project often categorizes types of crimes into two groups: index and non-index crimes. Index crimes include murders, criminal sexual assaults, aggravated assaults/batteries, burglaries, thefts, robberies, arson, and motor vehicle thefts. Non-index crimes include all other crimes, such as vandalism, weapons violations, public peace violations, and others. Sources, such as the report by Andrew Papachristos describe the spatial trends of crime in Chicago as well as the changes in terms of frequency over time of index crimes and the overall crime rate. According to Papachristos' report, Chicago's crime rates are similar to those of other U.S. cities. Crime rose in the 1960s, reached a high point in the 1990s, and since then it has steadily declined (Papachristos 4). Additionally, the report mentions that "especially socially and economically disadvantaged communities continue to have stubbornly high levels of crime" (Papachristos 6). Given the data that is used in this project, the team should be able to see this second trend in the rates of crime by community area in Chicago. Other sources found online present visualizations detailing the prevalence of types of crimes, such as theft versus battery. They also present time series plots of crimes over 17 years. An additional point of study was the trends of a particular crime, such as sexual assault, over a single year (Laughlin; Mangipudi).

3 PROPOSED WORK

3.1 Data Cleaning:

We will need to clean all of the data sets that we plan on using. This will involve ensuring that there are no missing or na values in the data sets and dropping unnecessary columns or empty rows from the data set. We will need to ensure that the data is in the correct format for use with other tools.

3.2 Data Preprocessing:

To ensure that our data is easy to use with the variety of tools that we intend to use on this project, we will need to spend a fair amount of time on data preprocessing. We will begin by calculating various metrics on the data, such as average crime rate, change in crime rate, arrest rate, and so on. Calculating this meta-data will be useful when we are creating visualizations of the data. Because a lot of the data is text-based, we will explore different methods

of tokenization that will allow us to use the text-based data with various machine learning tools.

3.3 Data Integration:

We are considering using weather data and census data to help us find potential relationships between the data sets. Our hope is to see some sort of relationship between the temperature and the crime rate. We will also attempt to see if there is a relationship between the average income of a neighborhood and the type of crime that is typical to the area.

3.4 Visualization:

Visualizing our data will be integral to finding interesting patterns and correlations in the data set. We will be looking for interesting trends in the crime rate in Chicago. Plotting the change in crime-rate over time will help us to identify specific crimes or specific locations that we should investigate further. We will be using various python libraries such as folium and seaborn to create useful visualizations that we can use to spot trends in the data. Folium will help us to easily heat map the data, allowing us to see which neighborhoods are worth looking into. This will help us to better understand how a particular neighborhood can get progressively worse over time.

4 DATA SET:

The main dataset comes from Chicago's online data catalog. The dataset lists roughly 6.54 million crime reports from 2001 to present. The columns that are most important to our project include the date when the crime occured, the Illinois Uniform Crime Reporting code, the ID of the community, a short description of the crime, a attribute indicating whether an arrest happened, the police district, and the geographical coordinates. There are other columns in the dataset, such as case number of the incident, the FBI code, or the ward ID, but we will unlikely utilize them. This dataset is about 1.5 Gigabytes on disk, which will pose some challenges for data manipulation. From initial data analysis, it looks like some of the data will need to be preprocessed to clean up invalid dates and missing data.

In addition to the main dataset of crime reports in Chicago, there are several other interesting datasets available through the city's data catalog. One of the datasets available describes various socioeconomic indicators of the different communities in Chicago. For example, it lists the percent of housing that is overcrowded, percent of households living below the federal poverty line, percent of people aged 16 and older that are unemployed, per capita income, and percent of persons over the age of 25 without a high school diploma. Given these community based indicators, it might be possible to find a correlation between these values and the rates of crime in the region. Other interesting datasets available on the catalog provide information on affordable housing and grocery stores in Chicago's communities. The catalog also publishes public health information such as the rates of elevated blood lead levels in children aged 0-6 years old as well as data on births and pregnancy care by community area. Like the socioeconomic indicators, these community statistics could provide less obvious links to crime rates.

5 EVALUATION METHODS:

Firstly, we hope to arrive at comparable results to previous studies of Chicago's crime. Secondly, we want to go further than previous studies and find novel trends between crime and other social and public factors. This knowledge could potentially influence Chicago's crime reduction strategies. For example, if we found that affordable housing was correlated with lower rates of crime that would be valuable knowledge for the Chicago city planners.

6 TOOLS:

6.1 Python:

The majority of our programming will be done in python. Python has a lot of useful tools that will help us work with this large data set. Some of the libraries we will be using in Python are Pandas, Numpy, Matplotlib, Folium, and Scipy.

6.2 Jupyter NoteBooks:

Jupyter notebooks are easy to use and great for exploring ideas quickly with python. Most of our work will likely involve jupyter notebooks in some way.

6.3 Pandas:

We will be using pandas to handle most of the data manipulation such as cleaning and structurization. Pandas will allow us to manipulate the data so that we can create better, more interesting visualizations.

6.4 Numpy:

It will be used for scientific computation and working with arrays.

6.5 Folium:

Used for mapping instances in the data set. We will use folium to heatmap different types of crime. This will help us to visualize the data and spot potential patterns to investigate.

6.6 Matplotlib:

This package will help us visualize the data after being processed. It will help us generate plots, histograms, bar charts, scatterplots, etc., with just few lines of code.

6.7 Scipy:

This library will be used to calculate statistics of our data. It will facilitate to calculate averages, means, medians, modes, z-scores and more statistical values.

6.8 RapidMiner:

Aside from coding in python, we will be using RapidMiner software to get predictive analytics and statistical modeling.

7 MILESTONES:

7.1 Milestone 1:

Have the data preprocessed - March 9th.

7.2 Milestone 2:

Create visualizations - March 16th. Augment crime data with another data set or sets - March 16th.

7.3 Milestone 3:

Create heat maps based upon districts - March 23rd.

7.4 Milestone 4:

Dive deep into the ’Weird’ data - April 6th.

7.5 Milestone 5:

Implement AI eg. (clustering, neural network, knn or random forest) - April 20th.

8 SUMMARY OF PEER REVIEW SESSION:

During the presentations we found that some of our peers were using the same data sets. Even though both of us will be applying the same data mining methods, their goal is to map the data to specific regions and present it to the Chicago Police Department to prevent further crimes. Our goal will be more focused on finding interesting trends or outliers in terms of overall crime rates or certain types of crimes, such as homicides.

9 WORKS CITED:

Mangipudi, Vivek. ANALYSIS OF CRIMES IN CHICAGO 2001 - 2017. 28 July 2017, rstudio-pubs-static.s3.amazonaws.com/294927_b602318d06b74e4cb2e6be336522e94e.html

Papachristos, Andrew V. fi48 YEARS OF CRIME IN CHICAGO: A Descriptive Analysis of Serious Crime Trends from 1965 to 2013.fi Yale ISPS, vol. 13, no. 023, 9 Dec. 2013, pp. 1fi?!20., isps.yale.edu/sites/default/files/publication/2013/12/48yearsofcrime_final_ispsworkingpaper023.pdf

Laughlin, Greg. fiCrime Over Time: Visualizing Crime Data in Chicago.fi Socrata, 3 June 2014, socrata.com/blog/crime-time-visualizing-crime-data-chicago/.